



2011-09

System Self-Awareness and Related Methods for Improving the Use and Understanding of Data within DoD

Zhao, Ying



Calhoun is a project of the Dudley Knox Library at NPS, furthering the precepts and goals of open government and government transparency. All information contained herein has been approved for release by the NPS Public Affairs Officer.

**Dudley Knox Library / Naval Postgraduate School
411 Dyer Road / 1 University Circle
Monterey, California USA 93943**

Data sources for Department of Defense (DoD) applications often include disparate real-time sensor and archival sources with multiple dimensions, very high delivery rates, and very large volumes over time. Analysts, at present, must manually comb through immense volumes of multisource, multi-classification-level intelligence data to find previously unknown and undiscovered patterns, associations, relationships, trends, and anomalies. In this article, the authors explain how their research and methodology can be applied to a set of data, leveraging automatic and intelligent tools to facilitate the decision maker's understanding of voluminous, dynamic data and to recognize hidden yet critical links and patterns derived from dynamic, ongoing DoD data. Their efforts could facilitate real-time awareness, reduce the workload of decision makers, and make a profound impact on the long-term success of DoD knowledge management. While their focus remains on DoD sources, this methodology can be applied to any data source.

Key words

data analysis, discursive space, entity extraction, learning agent, lexical link analysis, semantic networks, social networks, system self-awareness, theme extraction, unstructured data

System Self-Awareness and Related Methods for Improving the Use and Understanding of Data within DoD

YING ZHAO, DOUGLAS J. MACKINNON, AND
SHELLEY P. GALLUP

Distributed Information Systems
Experimentation (DISE) Group,
Naval Postgraduate School

INTRODUCTION

Data sources for DoD applications often include disparate real-time sensor and archival sources with multiple dimensions, as well as very high rates and very large volumes. The data can include structured data that are of traditional forms (for example, stored in relational databases, Excel, or XML files with well-defined labels with meta-data). Unstructured data can also be included, such as free text, Word, PDF, PowerPoint, and emails. A large percentage of data are unstructured. It remains a daunting task to retain logical integrity of the separate data sources, and support multiple parallel and asynchronous functions of storage, analysis,

search, and retrieval of these data sources. Analysts, at present, must manually comb through immense volumes of multisource, multiclassification-level intelligence data to find previously unknown and undiscovered patterns, associations, relationships, trends, and anomalies. This effort can be tedious and slow, and can depend largely upon the individual analyst's experience. Without accurate and timely front-end analysis, link analysis and inferences about future trends are not possible. The authors' methodology cross-examines and considers all of the data to create a full picture, and thus gain improved system self-awareness. Analysts and the intelligence community could benefit from automated, scalable, and robust tools and methods to analyze such large data sets quickly and thoroughly, to create and sustain situational awareness in real time.

For example, as a result of a recent near-disaster in the Christmas Day attack on a U.S.-bound airliner, the initial analysis from the White House indicated that there is a need for dramatic changes within the U.S. intelligence community to improve its information gathering, dissemination, and correlation (White House Report 2010); specifically, it states:

- U.S. counter-terror agencies lack a "Google-like" search capability. Google and other common Internet search engines routinely offer alternative spellings for searches, particularly with names. For example, in this attempted attack, Abdulmutallab had been flagged beforehand by U.S. embassy staff in Nigeria, but not under his full name. There lacks a particular standard for name-checks once a U.S. Visa is granted.
- The problem appears to be more about "connecting the dots" rather than a lack of information sharing after 9-11. For example, the U.S. government at the CIA and the National Counterterrorism Center (NCTC) had sufficient information prior to the attempted December 25 attack. The information was available to analysts, but fragmentary and embedded in a large volume of other data.
- Information technology within the counterterrorist community did not sufficiently enable the correlation of all-source data. There was a delayed dissemination of the intelligence reports to *all-source* analysts.

The authors' methodology, efforts, and proven capabilities to date might have been used to resolve

many of these concerns. Their research applies a set of data-driven automation systems to help DoD researchers and decision makers recognize important links and patterns derived from dynamic, ongoing large volumes of DoD data. The resulting systems could facilitate real-time awareness, reduce the workload of decision makers, and make a profound impact on the long-term success of DoD knowledge management. While their focus remains on DoD sources, this methodology can be applied to any data source.

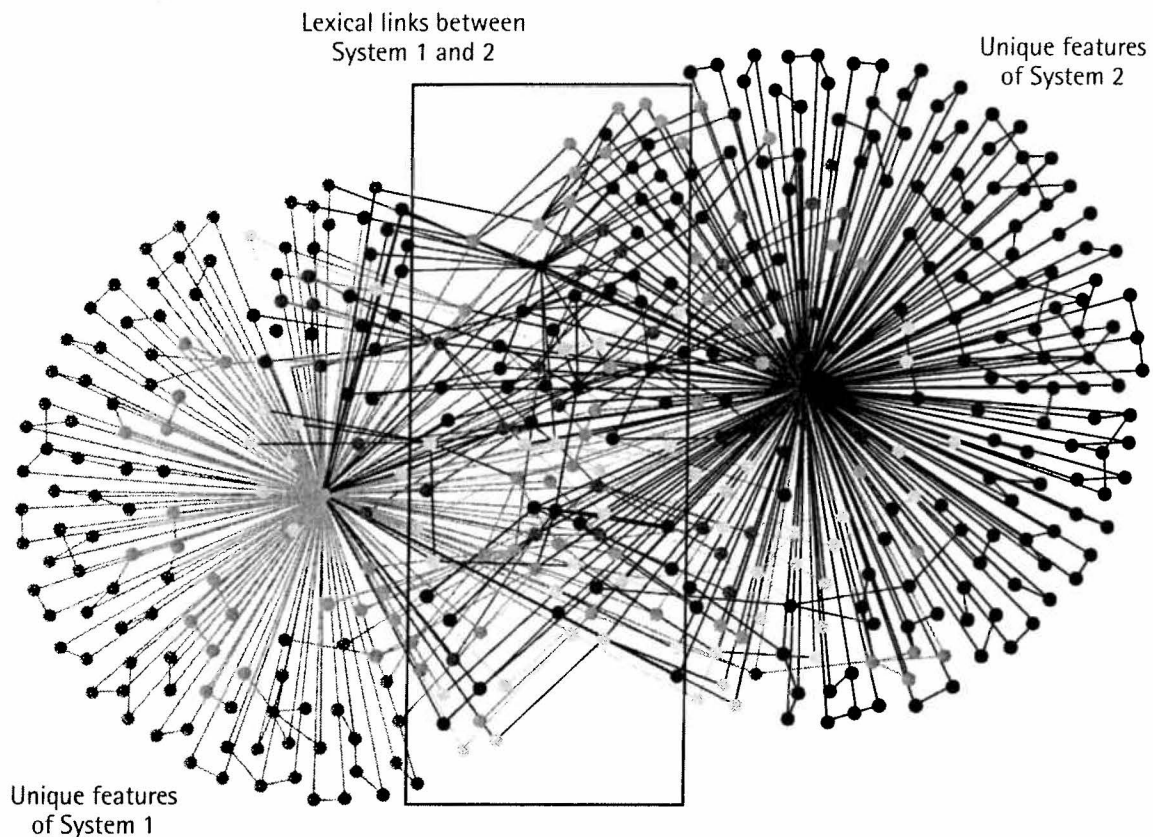
METHODOLOGY

Methods the authors have developed to date have proven useful in the analysis of very large data sets. There are several dimensions to their approach: 1) system self-awareness; 2) lexical link analysis; 3) agent learning; and 4) network analysis. The first represents a global view of an issue, and the other three refer to a set of specific methods and intelligent agent tools used to resolve analytic needs within these very large data sets.

System Self-Awareness

The authors borrow from notions of "awareness" and advance the term *self-awareness* of a complex system as the collective and integrated understanding of system capabilities, or *features*. A related term, "situational awareness," is used in military operations and carries with it a sense of immediacy and cognitive understanding of the war-fighting situation. They define *awareness* as the cognitive interface between decision makers and a complex system, expressed in a range of terms or "features," or specific vocabulary or "lexicon" to describe attributes and the surrounding environment in which the data are embedded.

Traditional data mining includes analytic tools that may be applied to both structured and unstructured data to confirm previously determined patterns, or to discover patterns that are yet unknown. Text mining is the application of data mining to unstructured or less-structured text files. Text mining represents an emerging field with a wide range of software implementing innovative visualization and navigation techniques. These techniques graphically represent networks of documentation that are related conceptually. Visualization of these relationships enables concept discovery, automated classification, and understandable categorization of unstructured documents by human decision makers.

FIGURE 1 Comparing two systems using LLA

©2011, ASQ

Lexical Link Analysis

Lexical analysis is a form of text mining in which word meanings represented in lexical terms (for example, word pairs) are developed from context. Link analysis, a subset of network analysis, explores associations between objects and reveals crucial, and sometimes surprising, relationships. Lexical link analysis (LLA) is an extension of lexical analysis combined with link analysis, and employs enabled agent learning technology. The detailed steps of LLA processing include applying collaborative learning agents (CLA) and generating visualizations, including a lexical network visualization via AutoMap, radar visualization, and matrix visualization (Zhao et al. 2010). Automation is achieved via computer program or software agent(s) to perform this task frequently—and in near real time.

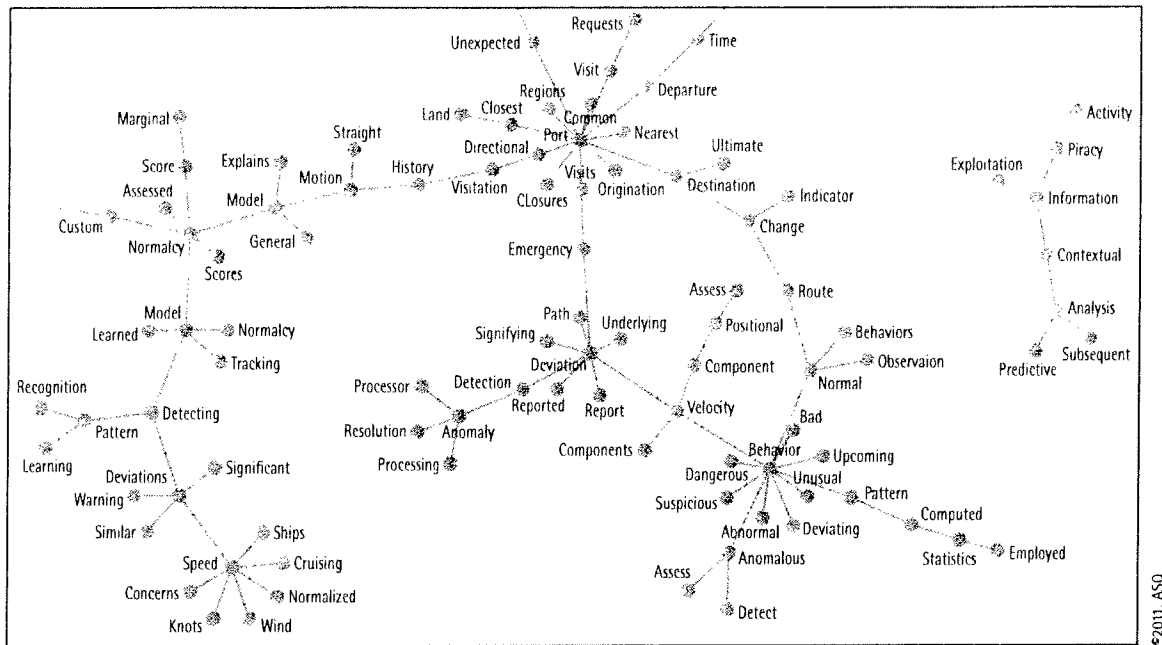
Link analysis “discovers” and displays a network of word pairs. These word pair networks are characterized by one-, two-, or three-word themes. The weight

of each theme is determined by its frequency of occurrence. Figure 1 shows a visualization of lexical links for Systems 1 and 2 of two competing systems, which are shown in the red box. Unlinked, outer vectors indicate unique system features.

Each node, or word hub, represents a system *feature*, and each color refers to the collection of lexicon (features) that describes a concept or theme. The overlapping area nodes are *lexical links*. What is unique here is that LLA constructs these linkages via intelligent agent technology using social network grouping methods.

Figure 2 shows a visualization of LLA with connected keywords or concepts extracted from the documents of Maritime Domain Awareness (MDA) technologies. Words are linked as word pairs that appear next to each other in the original documents. Different colors indicate different clusters of word groups. They were produced using a link analysis method—a social network grouping method (Girvan et al. 2001) where words are connected, as shown in

FIGURE 2 Features shown as word pairs



a single color, as if they are in a social community. A “hub” is formed around a word *centered* or connected with a list of other words (“fan-out” words) centered on other hub words. For instance, in Figure 3, the word “behavior” is centered with “suspicious, bad, dangerous, abnormal, usual, and anomalous” and so on, showing the ways to describe “behavior” in the MDA area.

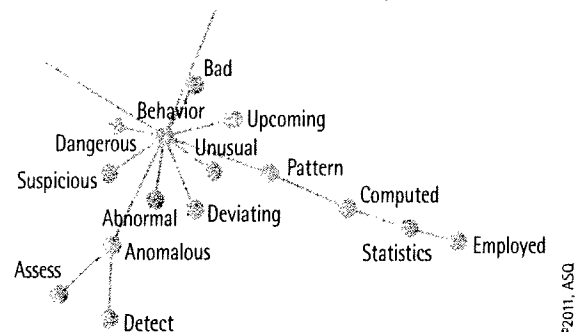
The following are the steps for performing an LLA:

- 1) Read each set of documents.
- 2) Select feature-like word pairs.
- 3) Apply a social network algorithm to group the word pairs into themes. A theme includes a collection of lexical word pairs connected to each other.
- 4) Compute a “weight” for a theme for the information of a time period, that is, how many word pairs belong to a theme for that time period.
- 5) Sort theme weights by time, and study the distributions of the themes by time.

Unsupervised Agent Learning

The authors have developed a computer-based learning agent to employ an unsupervised learning process

FIGURE 3 A Single word hub showing linkage (Detail of Figure 2)



that separates patterns and anomalies. They began at the Naval Postgraduate School (NPS) by using CLA (Zhou, Zhao, and Kotak 2009) and expanded to other tools. Results from these efforts arose from leveraging intelligent agent technology via an educational license with Quantum Intelligence, Inc. CLA is a computer-based learning agent, or agent collaboration, capable of ingesting and processing data sources.

Multiple agents can work collaboratively and in parallel. For this, the authors leveraged a large-scale parallel Linux cluster and cloud-computing

capability provided by the NPS High Performance Computing Center (HPC) to handle the data volume. The unsupervised agent learning is outlined by the following steps.

- 1) Index each set of documents separately and in parallel using multiple learning agents. The authors set up a cluster using Linux servers that are capable of accommodating hundreds to thousands of agents to process large volumes of the reports in the NPS Secure Technology Battle Laboratory (STBL) environment.
- 2) Apply context lists for entity extraction. Using word juxtaposition, context lists are provided initially to specify the contexts for who (people), where (location), and what (action). One uses context words in each category to:
 - Extract people using context words “commander, financier, brother, son, named, assistant,” and so on
 - Extract locations using context words “village, province, area,” and so on
 - Extract actions or events, extract verbs that are close to people, for example, “killed, injured, met, traveled,” and so on
 - Extracted word pairs are linked in the same events (for example, in the same documents or in the same paragraphs)
- 3) Generate social networks based on entities extracted. The relation types are people-to-people, location-to-location, action-to-action, people-to-location, people-to-action, and location-to-action. Each relationship is linked with a set of lexical terms that is discovered automatically from the data.
- 4) Generate semantic networks based on lexical links from the text documents that do not contain the entities extracted from the previous steps.
- 5) Apply visualization and network analysis highlighted next to analyze the extracted networks from Steps 1 to 4. Semantic networks combined with the people social networks will characterize the behavior, such as actions and events, of potential high-value targets.

Semantic and Social Network Analysis

Current research of social network analysis mostly focuses on using a simple, weighted, or arrowed link between two objects, such as people or organizations, regardless of the contents. The so-called study of centrality (Girvan 2002; Feldman 2007) has been a focal point for the social network structure study. Finding the *centrality* of a network lends insight into the various roles and groupings, such as the connectors (for example, mavens, leaders, bridges, isolated nodes), the clusters (and who is in them), the network core, and its periphery. The authors have been working toward three areas of innovations in the network analysis:

- Extract social networks based on the entity extraction
- Extract semantic networks based on the contents and word pairs using LLA
- Apply characteristics and centrality measures from the semantic networks and social networks to predict latent properties such as emerging leadership in the social networks. The characteristics are further categorized into themes and time-lined trends for prediction of future events.

Relations to Other Methods

LLA is related to a number of extant tools for text mining, including latent semantic analysis (LSA) (Dumais et al. 1988), key word analysis and tagging technology (Foltz 2002), and intelligence analysis ontology for cognitive assistants (Tecuci et al. 2007). What results from this process is a learning model—like ethnographic code book (Schensul et al. 1999). The authors’ extracted threads in social networks are related to entity extraction or named entity recognition (NER), which recognizes named entities such as people, organizations, locations, expressions of times, quantities, monetary values, and percentages in context. State-of-the-art NER systems (Hecking 2006; Army GIS 2010; MUC 1995-2005) for the English language produce near-human performance. For example, the best system scored 93 percent in accuracy, while human annotators scored 96 percent. However, these traditional linguistics-based models and supervised learning NERs are not feasible for real-life HUMINT

reports. An entity extraction system (Nadeau, Turney, and Matwin 2006) is an unsupervised learning system and does not need training data to start. Therefore, it is much faster and capable of ingesting much more data than many other methods. It has a reported 77.71 percent accuracy compared with a supervised annotation of 93 percent. A different approach, such as the DyNet tool suite (Carley 2007), uses dynamic network analysis tools to process unstructured data. The AutoMap in the tool set provides a user-friendly interface to visualize social networks and compute various methods related to the dynamic network analysis. Speed and scalability is the problem of AutoMap, which was tested on small data sets.

LLA is unique in the ability to construct these linkages discovered via intelligent agents using social network grouping methods, thus revealing underlying themes found within unstructured and unstructured data. For example, it is possible to show how the word “computing” links to “cloud” and “stream,” and quickly reveal new trends that are discovered from dynamic, real-time data. When compared with static word ontology for matching meaning, such as WordNet (2011) (developed at Princeton University, a lexical dictionary of English terms and their relationships derived manually as a static database over a very period of time), the authors’ approach is dynamic, data driven, and domain specific. Their methods, if conducted frequently and automatically, can reveal trends of the central themes over time, thus providing much-needed situational awareness for DoD applications.

CASE STUDIES

The authors believe their methodology is applicable to many problems both within and outside of the DoD, especially in the following categories:

- Understanding and managing massive, dynamic data
- Analysis of social, cultural, and linguistic data
- Multidisciplinary approaches to assessing linguistic data sets
- Extraction and representation of the information in nontechnical, unstructured documents
- Effective analysis of massive, unreliable, and diverse data
- Analysis of significant societal events

The authors have applied LLA in diverse areas of research. Diverse projects are underway at the NPS in the Distributed Information Systems Experimentation (DISE) group, which range from defense analysis, to acquisition research, to naval recruiting using Facebook. The authors discuss each of these in the next section from the perspective of improved data analysis using LLA.

Defense Analysis

Background and research questions

Collected data in the area of intelligence analysis arrives from myriad sources that overwhelm analysts’ ability to perform the necessary deeper intelligence analyses that result in timely situational awareness. Combining, analyzing, and making sense of such data sources as HUMINT—Human Intelligence, GEOINT—Geospatial Intelligence, IMINT—Imagery Intelligence, MASINT—Measurement and Signature Intelligence, OSINT—Open Source Intelligence, SIGINT—Signals Intelligence, COMINT—Communications Intelligence, ELINT—Electronic Intelligence, and Special Signals can be challenging. The goal for this project was to reconstruct and analyze social networks from raw HUMINT reports and capture changes in network dynamics over time. These reports typically reside in the classified network (JWICS). Research questions are listed as follows for this project:

- How can threat networks from large-scale raw reports be extracted automatically, accurately, and quickly?
- How can social networks be visualized to display the dynamics and changes over time?
- How can leaders and influential personnel be identified in a threat network?
- Where are the high-value targets in the network?
- How can future threat events be predicted based on past data?

Results

The authors performed an initial feasibility study using a few months of data composed of 80 data sets (approximately 133 MB) and approximately 1500 reports. Each report represented a separate event. Improvised explosive device (IED) specific data, for example, include post-blast information, and after-action reports from

the Combined Explosives Exploitation Cell (CEXC) (Phillips 2003) are voluminous and arrive unstructured. These data are collected and stored in the combined data network exchange (CIDNE 2010), which is a primary reporting tool used in Iraq and Afghanistan. CIDNE records events from diversified sources such as target development, civil affairs, psychological operations, engagement, or indirect fires. Significant Activities (SIGACTS) are also included, and report on IED incidents and attacks on coalition forces. In this challenging area, the authors' efforts demonstrated their capability to reconstruct social networks⁴ of people, places, and events, as well as to reveal trends and perhaps predict future events. They enhanced the intelligence analyst's ability to perform data searches, semantic indexing, data mining, entity extraction, and link analysis to improve the understanding of threat operations and facilitate the implementation of counter-terrorism operations.

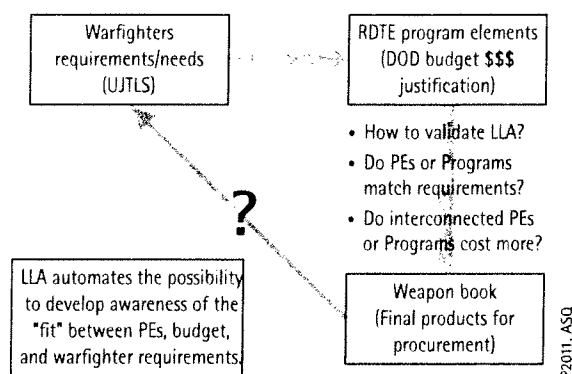
Acquisition Research

Background and research questions

DoD acquisition is an extremely complex system, comprising myriad stakeholders, processes, people, activities, and organizational structures. For example, there are three key processes in the DoD that must work in concert to deliver the capabilities required by the warfighters: the requirements process; the acquisition process; and the planning, programming, budget, and execution (PPBE) process.

Each process produces a large amount of data in an unstructured manner, for example, the warfighter requirements that are documented in universal joint activity lists (UJTL), joint capability areas (JCA), and urgent need statements (UNS). These requirements are considered using the Joint Capabilities Development System (JCIDS) process to perhaps become projects and programs, which should result in products such as weapon systems that meet the warfighters' needs. Program data are stored in the Defense Acquisition System (DAS). Programs are divided into Major DoD Acquisition Programs (MDAP) and Acquisition Category II (ACATII), among many others. Program elements (PEs) are the documents used to fund programs yearly through the congressional budget justification process. Data, here again, are too voluminous, unformatted, and unstructured

FIGURE 4 LLA seeks to link the business processes



to be easily digested and understood—even by a team of acquisition professionals. There is a critical need for *automation*, validation, and discovery to help reveal to decision makers and researchers the interrelationships within certain databases and inform their critical decisions.

The goal of this project was to data-mine massive documents for MDAPs, discover lexical links and drivers that impact program cost growth, and discover gaps with respect to warfighters' requirements.

The authors have attempted to develop and frame their research efforts around research questions in the following categories: conceptual, focused, theory development, and methodology.

Conceptual: Can the information that emerges from the acquisition process be used to produce overall awareness of the fit among programs, projects, and systems and of the needs for which they were intended? If a higher level of awareness is possible, will that enable system-level regulation of programs, projects, and systems for improvement of the acquisition system?

Focused: Based on the normal evolution of documentation and on the current data-based program information, can requirements (needs) be connected to system capabilities via automation of analysis? Can requirements gaps be revealed?

Theory development: What is the correlation between system interdependency (links/relationships) and development costs?

Methodology: Is it possible to use natural language and other documentation (roughly, unformatted data) to produce visualization of the internal constructs useful for management through LLA?

In precise terms, as shown in Figure 4, LLA seeks to link the business processes critical for DoD acquisition research, for example, linking requirements to DoD budget justification to final products. However, the authors observed that there were three important processes that seem fundamentally disconnected. Specifically, they were the congressional budgeting justification process (such as information contained within the PEs), the acquisition process (such as information in the MDAP and ACATII), and the warfighters' requirements (such as information in UNS and UJTLs). They further observed through interviews that they had never been analyzed nor compared together in a dynamic, holistic methodology that could keep pace with changes and reflect patterns of relationships.

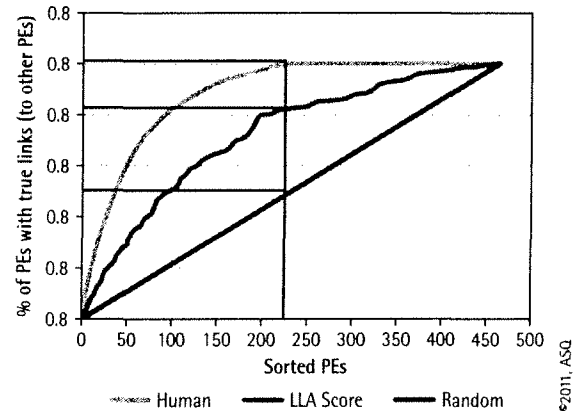
Results

To realize the potential of the LLA method, an important first step was to establish the validity of the method in the context of realistic, large-scale data sets. In the past year, the authors started to work on larger scale, open-source acquisition data sets. They obtained the research, development, test, and evaluation (RDT&E) congressional budget justification documents (for example, PEs from the DoD Comptroller website, <http://comptroller.defense.gov/defbudget/>). They also obtained program data, including MDAPs data and ACATII data, UJTLs data, and Weapon Books data from the DoD open-source websites and their OSD contacts. In total, they worked on part of the data of eight years (2002 to 2009) in the three services (Air Force, Army, and Navy)—about 5000 PDF documents.

The authors first applied LLA to extract the links based on PEs for the RDT&E congressional budget justification process. PEs were at the center of many documents because each PE listed all of the programs the PE funded and their costs for the one- and five-year projections. Specifically, they compared the trends of LLA with what human analysts had identified manually.

Figure 5 shows the accuracy of using the LLA method to predict the links between PEs. The x axis showed sorted PEs by three methods: random (green), LLA (red), and human (blue). The y axis showed the corresponding accumulative percentages of the PEs that were predicted correctly and that were linked to other PEs. The x - y curve was called a Gains Chart. As shown in Figure 5, there were about 225 PEs

FIGURE 5 Use of LLA scores to predict program element (PE) links: A Gains chart



that had at least one link to other PEs identified by human analysts (the blue line); 80 percent of them were predicted by the top 225 PEs sorted by the LLA scores (the red line). In other words, LLA was used to predict correctly 80 percent of the links identified by the human analysts.

NAVAL RECRUITING

Background and research questions

Facebook, YouTube, Twitter, and many other social networking sites offer virtual environments for meeting possible candidates that could fit service entry profiles. However, there are barriers preventing the Navy from fully capitalizing on the potential of these marketplaces. It is also difficult for recruiting decision makers to remain continuously aware of current marketplaces because social networks are relatively new platforms with overwhelming and dynamic information. Some recruiters also lack adequate technologies to monitor social networks or automatically match and link candidate profiles to service entry profiles to refine potential accessions.

The goal was to collect and match large-scale Facebook public fan and group profiles with Navy-enlisted and officer-rating documents that represent the requirements and needs for Navy Recruiting, to improve future Navy Recruiting and advertising efforts using this relatively new social platform. The research questions for this project are:

- How do current service entry profiles match with candidate profiles in the social networking environments?
- What methodologies facilitate matching profiles to potentially inform Navy Recruiting marketing decisions?
- Are there sufficient matches to make marketing in this environment cost effective?

Collecting samples of Facebook data is critical for the research in this project. Facebook collects rich information; for example, users can enter information on their background (such as high school, hometown), demographics (such as birthday, gender), interests, political views, and group affiliations, as well as on their cultural tastes (such as favorite books, movies, and music). Additionally, users can enter “friendship” relationships with users. Private information, however, is usually accessed only by friends and is not publicly available. The authors originally proposed to use a data set containing comprehensive Facebook profiles at several four-year American colleges that was made available for academic research (Lewis et al. 2008). The data set was later taken offline because of privacy concerns. As a result, they developed several different approaches to continue conducting the research.

Although access to private profiles is prohibited, Facebook contains volumes of data via public profiles. These profiles contain interesting information that list links and other pages and descriptions of the people themselves or other favored or “liked” sites. These pages are separated into categories such as books, movies, music, and TV shows. These fan pages sometimes show the preferences, interests, and behaviors of each person.

The authors started by comparing Facebook pages filled with publically available data and began comparing the text with data obtained from Navy Recruiting. The recruiting contained the needs and requirements for specific officer ratings. They further refined their search of Facebook profiles to only those exhibiting a “like” for colleges as well as the Naval Academy itself. This enabled them to locate more rapidly youth who might be a “match” for an officer billet in the U.S. Navy.

Results

As a result of this research, the authors also learned that they could collect and analyze “footprints” of where Facebook users had visited. These collected sites also indicate where Naval Recruiting might place

future advertisements. Ultimately, this effort resulted in a list of selected individuals who could become strong officer candidates in the U.S. Navy. The authors are in the process of applying the selection results periodically (for example, weekly) to continue to look for potential Navy officers.

Navy Chief of Information (CHINFO)

Background and questions

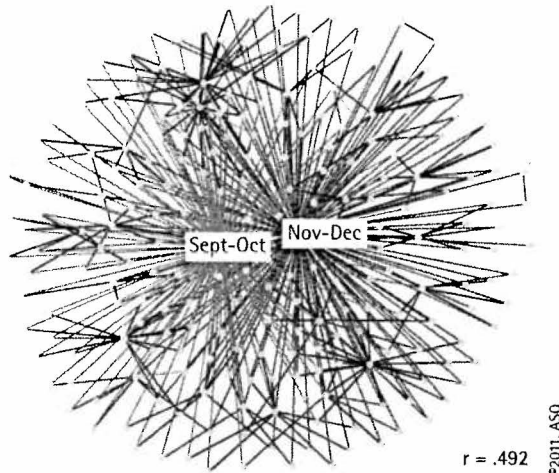
Discourse analysis has many meanings (Alvesson and Kärreman 2000; Fairhurst and Putnam 2004). The “big D” discourse refers to long-range, macrosystematic analysis, such as how democracy works and how it benefits a society as a whole. The “small d” discourse is local, situational analysis to ask a question, such as how discourse is being used in a particular context, and what types of strategies yield successful discourse. According to Fischer (2006), the discursive space is where the discourse takes place and can indicate how it has shaped the influence of stakeholders in the space. Using this approach, the authors again leveraged LLA to determine how strategic communications of CHINFO proliferate through various open sources. The case study analyzed involved the 2006 U.S. Coast Guard Live Fire case, when the Coast Guard planned a live fire training program in the Great Lakes area in Michigan. The program failed in the end because of public opposition. The data included a four-month public discourse recorded of about 980 public comments and 200 pages of public meeting transcripts. A manual qualitative analysis of themes was done by human analysts previously and compared with the results from LLA. The research questions were:

- How can one use the LLA to understand the dynamics within a given discursive space?
- What can the structural properties of the discourse reveal about who is participating and how they are shaping the space?
- What are the discursive strategies being used to shape the discourse space?

Results

The authors applied LLA by first analyzing and linking all associated comments, and then generating semantic networks over time by stakeholder groups. This effort,

FIGURE 6 Lexical link display for the discourse space for the Coast Guard



conducted over time, enabled them to analyze structural changes in the discourse space using quadratic assignment procedure (QAP) correlations (Hubert and Schultz 1976) in the UCINET tool.

The content of this discourse included the live fire initiative during the entire public comment period from September to December 2006. Different stakeholder groups included Coast Guard, individuals/citizens, nonprofit organizations, private business, and local/state government. They ran a QAP correlation on the different stakeholders and were able to create a highly detailed graph of the overlapping LLA images. By examining these correlations, the authors obtained a picture of the extent to which these stakeholder groups were aligned with one another, because of commonly used terms. Since the Coast Guard was the initiator, it was interesting to see how they correlated with the certain stakeholder groups. They overlapped with the others in the 10 percent range; that is, the QAP correlation was about 0.1.

Figure 6 shows an image of the Coast Guard's discourse strategy between time period 1 (September and October) and time period 2 (November and December). As one can see, there was a very high level of consistency in their own discourse with the QAP correlation of 0.49. The qualitative analysis done by humans validated that this came from a reliance of scripts for many of their interactions with other stakeholders. The high correlation of the Coast Guard's discourse indicated the Coast Guard did

not adapt its communication strategy to the public outcry and therefore was forced to withdraw the live fire training program.

According to the dialogue theory (Braten 1984), discourse strategies can be monological, meaning stakeholders have a monopoly, excluding or "swallowing" any other perspective. Discourse strategies can also be dialogical, meaning organizational closure and conceptual agreement is reached among stakeholders, which allows for consciousness and needs to cross the boundary or shift the domain, breaking a monopoly or monoperspective. The authors noted that the Coast Guard did not realize its position in the global social network of the decision process. They could have engaged more dialogical strategy and the outcome might have been different.

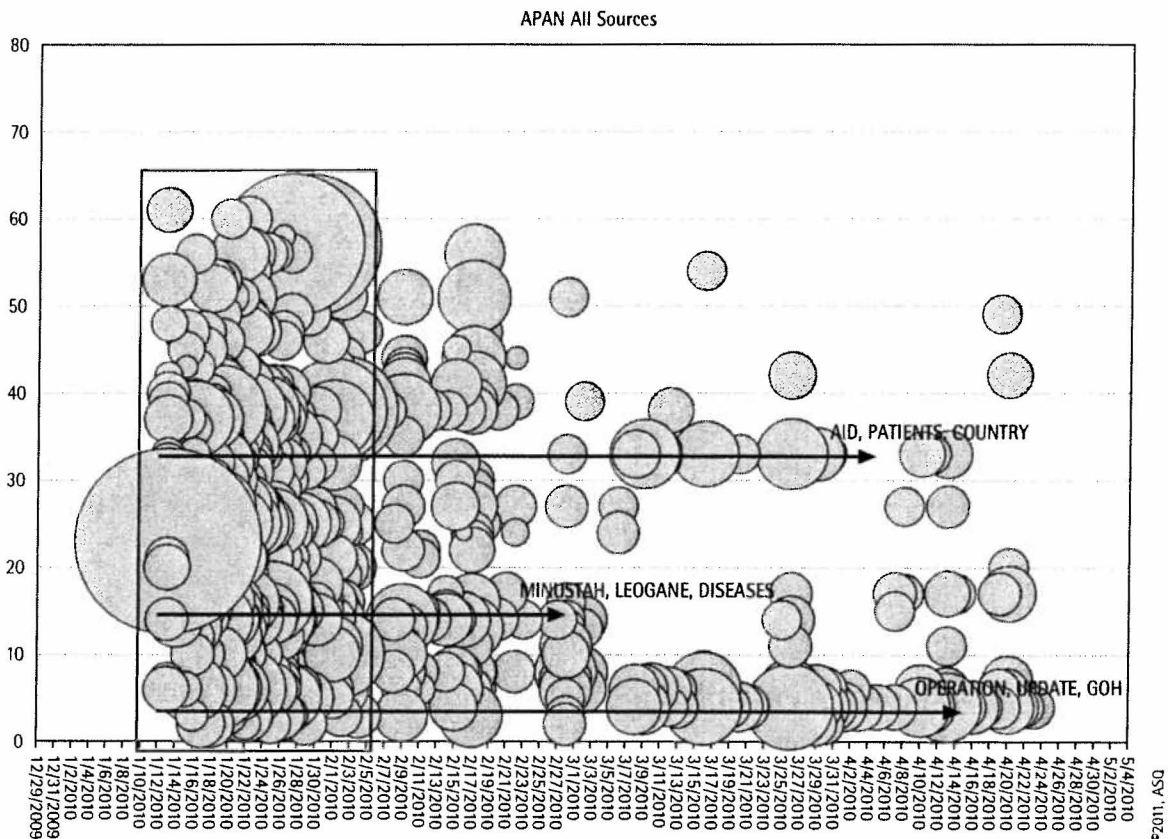
LLA allowed the authors to correlate and visualize a discourse and discursive strategies, and categorize them into various patterns. Their goal is to apply the method to understand any future discourses, evaluate which one is more effective and successful based on the objectives of the discourses, and try to influence or perhaps inform future discourses.

APAN Network and Haiti Operation Data Analysis

Background and research questions

As a matter of policy, the United States mobilizes diverse responses in aid, both internally and externally (to other countries), following natural disasters. In the aftermath of the Haiti earthquake, U.S. military and civil organizations provided rapid and extensive relief operations. Organizations involved in the operation collaborated with each other to negotiate roles, define limits, set business rules, and so on, to maximize efficiency and decrease response times. To determine that the expectations for effective and efficient operations had been attained would require in-depth analysis of collaboration, dialog, and resultant outcomes. Such analysis by manual means would require long and arduous effort. The authors proposed an automation of a portion of the analytic work, asking the questions:

- 1) What were the roles and relationships of these organizations?
- 2) How were the operations actually conducted?

FIGURE 7 APAN theme intensity as a function of day

For the lead military organization, there is an assumption that because the operational commander had resources and trained personnel, the expected outcome would be timely deployment of logistics, transport, security, and medicine. Also, there is an assumption among participating military organizations that due to their similar training, skills, and shared culture, their collaboration would result in higher effectiveness than civilian organizations. While anecdotal evidence supports this view, proving it as a fact is very difficult without tools such as LLA.

The challenge is to sift through the data that are collected in real-life events to create an overall picture of how various organizations (military and civil) actually collaborated, how their interactions were developed, and how their synergies were achieved. For example, how are the findings from real-life data different from participant expectations? How can these findings be used to improve the future operations?

Results

The authors used data samples and analysis from open sources during the Haiti humanitarian assistance operation. The sources included official PDF briefing documents, situation reports, forums, blogs, and open sources using web crawler technology, including Twitter, Facebook, and news-feed web sites.

They first collected approximately 2600 open-source web pages using a cluster of high-performance computer (HPC) nodes. The data were the news feed from Jan. 13, 2010, to Feb. 23, 2010, when the Haiti earthquake and international relief operations took place. The data were sorted by timestamps, domains, and organizations. The authors selected Twitter as a starting point since various organizations such as SOUTHCOM and USAID used Twitter to handle the situations that required real-time information gathering and dissemination in the Haiti relief efforts. Inrelief.org was a new site hosted by NPS that was of interest to the authors as to how it worked in a real-life event.

US SOUTHERN COMMAND had engaged the HAITI HA/DR Community of Interest (COI) on the All Partners Access Network (APAN) during the Haiti crisis. The APAN data were captured from a MSSQL database. The sources were:

- Official documents and briefings in PDF: approximately 167 PDF file attachments related to HAITI HA/DR from Jan. 13, 2010, to May 26, 2010
- SITREP: Approximately 150 situation report documents
- Forum: Approximately 1173 posts from Jan. 13, 2010, to June 3, 2010
- Blogs: Approximately 3900 blog messages

The authors' goal was to show initially that lexical links can be utilities and measures for trends of interagency operations. They collected the following lexical link measures.

- Number of nodes or word hubs and features
- Number of links between the word hubs
- Number of domains or organizations
- Number of cross-domains, for example, Twitter/SOUTHCOM is a cross-domain when SOUTHCOM used Twitter to communicate
- Synergy index: Defined as the number of word hubs between two organizations divided by the total number of word hubs from the two organizations.

Synergy is normally a measure obtained using the traditional QAP used in social sciences and social network analysis. Synergy defined here, however, is calculated from lexical links instead of social network links, which are commonly used.

The authors seek to compare the themes discovered from the sources and to see how organizations were collaborating during the operation. Figure 7 shows an example of a "bubble plot" of the timeline of all of the themes. In Figure 7, the x-axis represents the day and the y-axis represents the theme identification number. A theme bubble is centered on a number for a particular day, and the bubble size represents the number of lexical terms that describe that theme. A large-size bubble means there is a lot of information (that is, lexical terms) on a given day. The first two weeks (in the green box) depict all of the major themes

covered. Some themes continued, such as "Operation, Update, GOH," and some dominated only later, such as "Minustah, Leogane, Diseases" and "Aid, Patients, Country."

The authors found that by using samples of open source information for the interagency collaboration of Haiti earthquake relief operations, the number of overlapping lexical links can be used to measure synergy between two collaborating organizations. The synergy among the organizations in the Haiti relief operations was low in the beginning and became higher as time went on. They also found that social media such as Twitter, Facebook, and Google provided critical capabilities for fast information gathering and dissemination for military entities such as SOUTHCOM. LLA provides a method to look at the interagency collaboration directly from the real-life communications and documentations.

CONCLUSION

The authors' efforts and research indicate that their methodology provides analysts the means to analyze very large amounts of available data—in a variety of contexts. Central themes are depicted graphically among a set of automated and intelligent tools. Specific searches can be developed for individual words or phrases, yet it is the natural, unsupervised discovery, allowing the data to reveal its themes and linkages, that has revealed interesting and surprising findings.

The authors seek to move into real-time analysis of frequently changing databases and depict them dynamically in three dimensions. They believe they are now poised to reveal a change in the fundamental methods used to understand and depict large sets of data, thus allowing human decision makers heretofore unreachable insight into their disparate, multisource, structured and unstructured, voluminous, and dynamic data.

REFERENCES

- Alvesson, M., and D. Kärreman. 2000. Varieties of discourse. On the study of organizations through discourse analysis. *Human Relations* 53, no. 9:1125-1149.
- Army GIS. 2010. Available at: http://www.agc.army.mil/gis/gisint_body2.html.
- Braten, S. 1984. The third position—Beyond artificial and AutoPoietic reduction. *Kybernetes* 13:157-163.

Carley, K. M. 2007. Improved data extraction and assessment for dynamic network analysis, Special Report 66. U.S. Army Research Institute Program in Basic Research. Available at: http://www.hqda.army.mil/ari/pdf/S_66.pdf—699.764 KB—29 JUL 2008.

CIDNE. 2010. Available at: <http://www.issinc.com/solutions/cidne.html>.

Dumais, S. T. et al. 1988. Using latent semantic analysis to improve information retrieval. In *Proceedings of CHI'88: Conference on Human Factors in Computing*, 281-285.

Feldman, R. 2007. Link analysis and text mining: Current state of the art and applications for counter terrorism. In NATO Advanced Study Institute on Mining Massive Data Sets for Security.

Fischer, K. ed. 2006. *Approaches to discourse particles*. Amsterdam and Boston: Elsevier.

Foltz, P. W. 2002. Quantitative cognitive models of text and discourse processing. In *The Handbook of Discourse Processes*. Mahwah, NJ: Lawrence Erlbaum Publishing.

Gerber, C. 2005. Smart searching: New technology is helping defense intelligence analysts sort through huge volumes of data. *Military Information Technology* 9, no. 9.

Girvan, M., and M. E. J. Newman. 2002. Community structure in social and biological networks. Available at: <http://www.pnas.org/content/99/12/7821.full.pdf>.

Hecking, M. 2006. The state of the art and the state of the practice: Content analysis of HUMINT reports. In *Proceedings from the Command and Control Research and Technology Symposium (CCRTS)*.

Lewis, K., J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis. 2008. Tastes, ties, and time (T3): A new social network dataset using Facebook. *com. Social Networks* 30:330-342.

MUC. 1995-2005. Message Understanding Conference. Available at: http://www-nlpir.nist.gov/related_projects/muc/.

Nadeau, D., P. D. Turney, and S. Matwin. 2006. Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. Canadian Conference on Artificial Intelligence, Québec City, Québec, Canada. June 7.

Pajek. 2011. Available at: <http://vlado.fmf.uni-lj.si/pub/networks/pajek/>.

Phillips, S. 2003. The birth of the combined explosives exploitation cell. Available at: <http://www.smallwarsjournal.com>.

Schensul, J. J., S. L. Schensul, and M. D. LeCompte. 1999. *Essential ethnographic methods: Observations, interviews and questionnaires*. Lanham, MD: Rowman Altamira.

Tecuci, G., M. Boicu, D. Marcu, C. Boicu, M. Barbulescu, C. Ayers, and D. Cammons. 2007. Cognitive assistants for analysts. In Joint Publication of the National Intelligence University, Office of the Director of National Intelligence, and U.S. Army War College Center for Strategic Leadership.

WordNet. 2011. Available at: <http://wordnet.princeton.edu/>.

White House Report. 2010. Available at: http://www.whitehouse.gov/sites/default/files/summary_of_wh_review_12-25-09.pdf.

Zhou, C., Y. Zhao, and C. Kotak. 2009. The collaborative learning agent (CLA) in Trident Warrior 08 exercise. *International Conference on Knowledge Discovery and Information Retrieval, Madeira Portugal (KDIR)*, October.

Zhao, Y., D. J. MacKinnon, S. P. Gallup, and C. Zhou. 2010. Maritime domain awareness via agent learning and collaboration. In *Proceedings of the 15th ICCRTS, International Command and Control, Research and Technology Symposium*, Santa Monica, California, June 22-24.

Zhao, Y., S. P. Gallup, and D. J. MacKinnon. 2010. Towards real-time program-awareness via lexical link analysis. In *Proceedings of the 7th Annual Acquisition Research Symposium*, Monterey, California, May 11-13.

BIOGRAPHIES

Ying Zhao joined the Naval Postgraduate School (NPS) in June 2009 as a research associate professor in the Information Sciences Department and Distributed Information Systems Experimentation (DISE) Group. Her research interests include theories and methodologies of quantum intelligence, data/text/information mining, pattern recognition, lexical link analysis, social network analysis, agent learning, search, and visualization. She has also been working on large-scale applications of these methodologies to DoD interests. Before joining NPS, Zhao co-founded Quantum Intelligence (QI), Inc. She was the principal investigator for six small business innovation research and federal awarded projects for QI from 2002 to 2008. She is the co-author of two pending patents, and received a special commendation for Outstanding Achievement in the Development of Bio-Spice Technology from Defense Advanced Research Program Agency (DARPA) in 2005. Zhao can be reached at yzhao@nps.edu.

Doug MacKinnon is a research associate professor at the Naval Postgraduate School (NPS). He is the chief of staff of the Distributed Information Systems Experimentation (DISE) research group. Since 2009, he has become involved with data mining research and its effect on acquisition and knowledge management. In 2008, he led an NPS research team to assess new MDA, spiral-1 technologies being fielded by PEO C4I developing original decision matrix structures and metrics structures to leverage the new technology. He also participated in the assessment of the tasking, planning, exploitation, and dissemination (TPED) process during the Empire Challenge 2008 (EC08) and leads the assessment this year. He holds a doctorate and an engineer's degree from Stanford University. He holds two master of science degrees, one in information technology management and the other in operations research from the Naval Postgraduate School (NPS). His career includes nine years on NPS faculty, teaching graduate operations research courses in probability and statistics, wargaming, and simulation.

Shelley Gallup is a retired surface warfare officer, an associate research professor at the Naval Postgraduate School's Department of Information Sciences, and director of Distributed Information and Systems Experimentation (DISE). He has a varied science, engineering, and analysis background including microbiology, biochemistry, space systems, international relations, strategy and policy, and systems analysis. He returned to academia after retiring from Naval service in 1994, and received his doctorate in engineering management from Old Dominion University in 1998. Gallup joined NPS in 1999, bringing his background in systems analysis, naval operations, military systems, and experiment methods first to the Fleet Battle Experiment project (1999 to 2002) then to the FORCEnet experimentation in the Trident Warrior series of experiments (2003 to 2008). His particular expertise has been development of complex experimentation project management, knowledge management, and methods to be used in military utility assessments.